Brittany Hayes. Detecting Lexically Cohesive & Temporally Bounded Tweet Sessions on Twitter Timelines. A Master's paper for the M.S. in I.S. degree. April, 2014. 59 pages. Advisor: Stephanie W. Haas

This exploratory study examines the concept of the "tweet session"—instances where a user of the microblogging site Twitter posts two or more related tweets in a short period of time. The study outlines a set of characteristics that aid in the detection of topically cohesive units. Four of these properties are external to the tweet text: a 24-hour time frame, inclusion of at least two tweets, inclusion of originally authored tweets and the exclusion of replies. Four additional properties were derived from the natural language processing and information retrieval literature: lexical cohesion based on unigram and character bigram feature representations, conjunction use, signals of continuation, and anaphora resolution.

A sample of 220 user timelines was analyzed to detect series of tweets meeting the definition of a session as conceptualized in the study. 93.6% of timelines included at least one technical tweet session. Lexical cohesion as determined by cosine similarity retrieved the most sessions: 815 technical sessions when unigrams were used as the unit of tokenization, and 1391 technical sessions when character bigrams were used. The majority of users engaged in tweet sessions exceeding 140 characters (the Twitter-imposed limit for a single tweet); however, when unigrams were used in the feature representation approximately 47% of timelines had tweet sessions of less than 140 characters on average. This research shows that tweet sessions exist and can be detected by computational means.

Headings:

Microblogs

Natural Language Processing (Computer Science)

Text Mining (Information Retrieval)

DETECTING LEXICALLY COHESIVE & TEMPORALLY BOUNDED TWEET SESSIONS ON TWITTER TIMELINES

by
Brittany Hayes

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April, 2014

Approved by:

_____

Stephanie W. Haas

Table of Contents

List of Tables

List of Figures

Introduction

With the rise of social media, individuals can easily and immediately share their thoughts and opinions. Microblogging has emerged as one of the most popular modes of social networking for such purposes. This medium allows users to post brief updates of an instant nature, many times during a single day. First released in late 2006, Twitter is one such microblogging platform that allows its users to "tell the world what's happening in 140 characters or less."[1] The rapid adoption of this mode of communication has resulted in the growth of rich content, published by a diverse public in an up-to-the-second manner.

Twitter data is certainly "big data," and it has been deemed valuable to many groups—from politicians to non-profit organizations to large corporations—who seek to better understand the public's perception. Particularly in the case of for-profit businesses, Twitter can provide a wealth of information of use in sentiment analysis, which aims to detect the emotion expressed by an author in a text. Understanding how customers feel about one's latest product or business move can provide insight into future planning decisions, as well as giving companies the ability to act immediately if the situation warrants a brisk response. As Lim et al. (2013) wrote:

[1] Twitter: https://twitter.com/

> By amassing a large volume of timely feedback and opinions from diverse social media users and analyzing them using social media analytics, one can derive a wide range of social and business insights much needed for social policy formulation, customer relationship management, and product innovation…Advanced information extraction, topic identification, opinion mining, and time-series analysis techniques can be applied to traditional business information and the new BI 2.0 contents for various accounting, finance, and marketing applications, such as enterprise risk assessment and management, credit rating and analysis, corporate event analysis, stock and portfolio performance prediction, viral marketing analysis, and so on (2).

However, the previously mentioned 140-character limit to a Twitter post (or, "tweet") can make sentiment analysis a difficult task. As a single tweet must convey meaning in a small amount of space, it is inevitable that ambiguity will occur. On a basic level, ambiguity arises from a lack of context. While this problem cannot always be rectified, it can be mitigated in some cases. Particularly, this is likely to be possible through the identification of multiple tweets made by a user that discuss the same topic—what will be defined herein as a "tweet session." This exploratory study seeks to use natural language processing techniques to identify series of related tweets on user timelines. Using a sample of data collected from Twitter.com, the following research questions will be explored:

RQ1: What percentage of users engages in tweet sessions?

RQ2: How many sessions are detected using various techniques for detecting cohesion? How do different feature representations perform?

RQ3: On average, how long is a tweet session (in terms of textual length)?

RQ4: How influential are retweets in tweet sessions?

This research provides insight into how users overcome Twitter's constraints to express themselves in more than 140-characters, which could lead to novel ideas regarding how social media analytics can be improved. Additionally, it explores the

utility of various techniques that can be used to retrieve tweet sessions by computational means.

Literature Review

This review provides a brief overview of general studies of the social networking service, Twitter. The discussion proceeds to focus on discourse analysis and search session detection as means of segmenting information into cohesive units, presenting some of the methodologies utilized in both of these areas of research. The review ultimately considers the similarities and differences of these areas of research in relation to the current study.

Insight into Twitter

Understanding the medium of Twitter is an important step in approaching the study of the activity occurring on user timelines. Fortunately, there is a large body of literature available that can aid in this task. In the years since its inception, Twitter—and microblogging in general—has become a popular topic of research. Java et al. (2007) provided one of the earliest and most comprehensive looks at this social networking service; it is one of the seminal articles in the body of literature surrounding Twitter and microblogging. The researchers examined general trends, such as the increasing numbers of users and posts on the social networking site. The dataset of 1.3 million tweets used by the researchers was gathered from the site over a two-month span in 2007. The

researchers found that Twitter had grown steadily and globally. Additionally, Twitter usage falls into certain identified categories (daily chatter, conversations, sharing information, and reporting news), with any given user potentially using the service in different ways in relation to the numerous communities he may find himself a part of.

Though Java, et al. does not specifically provide a quantitative look at tweet length over the social network, it does make explicit reference to the 140-character limit enforced by the service. One study that did examine tweet length is Alis & Lim (2013). The researchers collected a corpus of 229 million tweets over a three-year period. They used this dataset to examine the changes in tweet length over this span of time, making a distinction between all tweets and "conversational utterances." The latter is the subset of tweets that make use of the @reply functionality. The study found that median tweet length fell from 10 words in 2009 to 8 words in 2012, and median utterance length fell from 8 words to 5 words in the same period. As this study showed, the brevity of tweets is a unique factor that has a major impact; it implies that short message length is a characteristic that is only getting more pronounced as time progresses.

Detecting Topical Shifts in Media

Though Twitter-related research abounds, it does not appear that many researchers to date have looked specifically at how topic shifts occur on a single user's microblog timeline. However, this has been extensively studied in other contexts. Situated in a multidisciplinary context spanning a range from natural language processing to information retrieval, several approaches to this problem have been proposed. Two

major areas that inform the present research—discourse analysis and segmentation, and identification of search sessions—are discussed below.

Discourse Analysis. Discourse—and therefore discourse analysis—is an ambiguous concept that scholars in various fields have defined and interpreted in a number of ways. In the introduction to their anthology of articles concerning discourse analysis, editors Hamilton, Schiffrin & Tannen (2001) noted that, though various conceptualizations of "discourse" abound, they all concern themselves with three main issues: "(1) anything beyond the sentence, (2) language use, and (3) a broader range of social practice that includes nonlinguistic and nonspecific instances of language" (1).

Linguistics scholars tend to focus on the first two of these categories. Unsurprisingly, a number of studies in the computational linguistics and natural language processing areas also take this view; these are of particular interest in informing the present study. Grosz and Sidner (1986) presented an early, influential theorization of discourse as it applied to the computational linguistics field. The scholars defined three elements that are found in any discourse: the linguistic structure, the intentional structure, and the attentional state. The linguistic structure is the division of a discourse into sequences of related utterances, each of which serves as a discourse segment. As they noted, each utterance fulfills a given role in the discourse segment; furthermore, each discourse segment fulfills a given role in the larger discourse. The concept of intentional structure dealt with the idea that any discourse must have an underlying purpose, which enables hearers and readers to recognize unique, coherent discourse segments (and the discourse as a whole). The last component—attentional state—was defined as the parts of

a discourse that will draw the focus of participants at a given point during the course of the discourse.

Uncovering discourses by computational means is a problem that many scholars have tried to solve. The first step to this discovery is the uncovering of the discourse's linguistic structure. Morris & Hirst (1991) noted: "A text or discourse is not just a set of sentences, each on some random topic. Rather, the sentences and phrases of any sensible text will each tend to be about the same things—that is, the text will have a quality of unity" (21). They argued that lexical cohesion—measured in terms of the semantic connections that exist between words—is an effective way to capture related segments, presenting an early approach for segmenting discourses by measuring lexical cohesion. The researchers used the term "lexical chains" to describe sequences of related words aligning to a unified topical unit of a text; lexical chains serve as a concrete manifestation of lexical cohesion.

In the study, Morris & Hirst used a sample of five texts from magazines. After manually identifying lexical chains in all of the texts, the researchers used these observations to formalize an algorithm for the detection of these lexically related segments. In their work, they use Roget's Thesaurus (1977) as a knowledge base in order to help identify the similarity of tokens in their texts. Some of their considerations included the type of thesaural relation existing between words, the level of transitivity of word relations, and the distance (in sentences) allowable between words in a chain. Some specific rules incorporated into their ultimate algorithm included: words would be considered related if they had at most one transitive link, and the number of sentences between related words could not exceed three (34).

Hearst (1994) is another study that focused on measuring lexical cohesion—here, defined as term repetition between sections of text—as a means of uncovering the structure of texts. In this study, Hearst introduced the highly cited algorithm, TextTiling. Unlike the previous approach by Morris & Hirst, Hearst utilized a highly automated series of operations to uncover cohesive segments of text. Her approach began with a tokenization of the text sample. Subsequently she measured similarity between features, and finally identified boundaries between subtopics of the text. In this study, cosine similarity was used as the metric to compare distance between sentences.

The above studies focused specifically on the closeness of semantic relations between words appearing in the texts under analysis. However, there are other features that have been considered as signals of cohesion in the literature. Of particular interest are form features. For instance, Hirschberg & Litman (1993) examined the use of cue phrases, defined as "words and phrases that directly signal the structure of a discourse" (501). Their sample consisted of spoken language, both in audio form and its corresponding transcribed text form. The researchers attempted to discover when cue phrases were used sententially (i.e. as an adverb that helps the hearer/reader interpret a given utterance) or in a discourse sense (i.e. signifying a digression from the present topic). The latter sense is useful in terms of segmenting discourses. In speech corpora, the researchers found that the use of intonational phrasing and pitch accent helped disambiguate these forms. In transcribed text corpora, orthographic symbols and part of speech tagging were features that could aid in disambiguation.

Similarly, Galley et al. (2003) considered form features in their analysis of spoken language texts. They discovered that combining content-based features with form cues

generally performed as well as or better than (and at a statistically significant level) existing approaches to segmentation on speech data. Some of the form cues that they considered included presence of particular phrases, silences, overlapping speech, and the introduction of a new speaker. While the combination of such features with lexical cohesion performed better than either approach alone (i.e., only lexical cohesion or only form cues), the researchers noted that lexical cohesion generally has a stronger influence in detecting topic breaks than the other features.

Given the diversity of sample types in the studies discussed, segmentation of discourses into cohesive, coherent chunks is clearly relevant for a number of different media types. As briefly outlined in Galley et al., some of these contexts include text, recorded speech and video; in spite of these differences, the ultimate goal for each is to be able to divide the language that makes up each into "topically related units" (562). Though the task is generally the same in each context, the literature reveals that the characteristics of each influence the methods for performing the segmentation.

For instance, as was discussed, studies such as Hirschberg & Litman and Galley et al. tackled the issue of segmentation within the realm of audio recordings. Specifically, Hirschberg & Litman used a corpus of multi-speaker utterances from a radio call-in program, as well as a keynote address given by a single speaker; Galley et al. used a corpus of recorded meeting minutes. Both of these studies were thus able to utilize a set of form features that are not available in traditional written text, or even transcribed texts. Even in comparing results between audio recordings and transcriptions of those recordings, as was done by Hirschberg & Litman, it becomes clear that certain features in

one medium that are not present in another can be particularly revelatory. For instance, in examining use of the word ‘now’ in the radio corpus, those researchers found that accent type, phrasal composition and phrasal position could all be used to disambiguate between the discourse form of ‘now’ and the sentential form of ‘now’ with high reliability. Their examination of cue disambiguation of the transcribed radio program corpus proved more limited. The only orthogonal feature for that dataset that could reliably inform this determination was related to phrasal position; if a punctuation mark or speaker name preceded the term ‘now,’ it was predicted to be in first position—and most instances where ‘now’ was in first position were discourse versions of the term.

As is clear from that comparison, written text requires a different set of features for effective discovery of underlying cohesion. However, many past studies have focused on relatively ‘traditional’ texts—those that are well formatted, and lengthier than a tweet. For instance, Hearst focused on long, expository text, specifically a popular science publication entitled *Stargazers*. Similarly, while the corpus of popular press articles used in Morris & Hirst was not extremely large—it consisted of a total of 183 sentences—a single document is still substantially longer than a tweet. Published articles are also much more likely to be more grammatical than a typical tweet on a user timeline.

Search Log Analysis & Search Sessions. Another relevant area of research that has been explored in the information retrieval community involves the examination of user interaction with search engines through query logs. This research has been performed in order to better understand user information needs and how search engines can fulfill them. There are a number of similarities between the structure of search

queries and tweets that can help inform the identification of the 'tweet session' in the context of microblogging. He et al. (2002) noted that search queries are very short, chronologically organized segments of text. Additionally, the goals of search session detection are similar to those of tweet session detection as it applies to this study. As He et al. noted, reconstruction of search sessions is meant to "group together search activities related to the same search topic and treat them as a whole during the process of identifying the search contexts" (729). In short, researchers working in this area attempt to group topically cohesive queries together, with the goal of better understanding users' intent.

Methodologically, researchers have approached the goal of uncovering search sessions in similar ways to those focusing on discourse analysis. One of the major approaches of these researchers is also to consider the lexical similarity of queries. For instance, Jansen et al. (2007) compared three different methods for session detection: IP address and cookie; IP address, cookie, and a temporal cutoff; and IP address, cookie, and context changes (864). The last case was the only one that considered the actual content of the query, mainly focusing on the repetition of the same terms occurring in queries from the same session. The researchers found that this method performed best.

There are, however, elements beyond the query text that are effective in session detection. Jones & Klinkner (2008) outlined four major categories of features, which they tested in their experiments. In addition to word and character edit features (i.e. query terms and characters in common between queries), the researchers noted that temporal features, query log sequence features, and web search features might also be useful.

These included qualities such as the number of words in common between queries, the timespan between queries, the log-likelihood ratio score (LLR) of a co-occurring query pair[2], and Prisma.[3] Ultimately, the best performing classifier in Jones & Klinkner took advantage of features from all four categories, achieving accuracy above 90%.

While the tweet session does not appear to have been studied prior to this work, several interconnected domains have tried to solve similar problems. The approaches set forth in the related literature can greatly aid in the understanding of the current task. Given the similarities and differences of Twitter analysis, discourse analysis, and search session detection with respect to the current study, a methodology incorporating elements gleaned from each is most appropriate.

[2] A statistical test which "indicat[es] that the pair of queries occur in sequence more than could be expected by chance" (Jones & Klinkner, 704)
[3] The "cosine distance between vectors derived from the first 50 search results for the query terms" (Jones & Klinkner, 705)

Methods

In this section, I formally define the concept of the tweet session, discussing each quality that signals tweet inclusion or exclusion in a session. I also provide an outline of the data collection method, including a brief discussion of the relevant Twitter Application Programming Interfaces (APIs) used. Lastly, this section details the ethical considerations guiding the research method.

Definition of Tweet Session

At the heart of this research is the concept of the 'tweet session.' In this study, a tweet session refers to a series of Twitter posts (tweets) on a single user's timeline, each of which are related to the same topic. More specifically, given the character limit enforced by the service, a tweet session is any instance where more than one tweet and/or more than 140 characters is posted about the same topic within 24 hours on the user's timeline.

The main constraints as inferred through this definition relate to: (1) time, (2) tweet quantity and length, (3) tweet authorship, (4) intended audience, and (5) cohesion. Table 1 presents an overview of criteria for tweet sessions related to the first four concepts.

Table 1: Criteria for Inclusion within a Session

Criterion	**Examples**
Temporal: First/Last tweet of session must occur within 24 hours of anchor (central) tweet of session	Candidate: Merry Christmas Everyone! [12/25/13 8:57] Got an IPad air for Christmas ! [12/25/13 10:00] NOT a candidate: Big Boi Speaks On OutKast Reunion (URL) [1/20/14 18:29] Andre 3000 And Big Boi May Drop Solo Albums (URL) [1/21/14 23:24]
Number of Tweets: Sessions must consist of two or more tweets	Candidate: I guess I'll watch this skins game [12/8/13 18:02] Such quality tackling... Proud to be a skins fan right now [12/8/13 18:33] NOT a candidate: I guess I'll watch this skins game [12/8/13 18:02]
Authored Tweets v. Retweets: Sessions must contain at least one tweet with authorship attributed to the user. Sessions cannot be made up entirely of retweets; however, retweets can be included in a session	Candidate: So #GOVCuomo doesn't think people that RESPECT #Human #LIFE should be in NY. Who the heck does he want in #NY. 1scary ind. something's wrong [1/20/14 5:20] RT @(USER): Someone got to @(USER). He dropped contemptuous use of ""right to life""(whoops) returns to ""anti-choice"" (URL) [1/20/14 5:21] NOT a candidate: RT @(USER): Cakes of the Poets: Derek Walcott is 84 today. This is the cake made for him at Ladera in Saint Lucia: (URL) [1/23/14 20:40] RT @(USER): Greater Antillean Bullfinch eating Derek Walcott's birthday pudding: #Faber (URL) [1/23/14 20:40]

Criterion	**Examples**
@Replies: Sessions and conversations are not equivalent, and therefore sessions cannot contain @replies	NOT a candidate: @(USER) I don't like you .. [1/23/14 03:35] @(USER) Jene doesn't like you ..[1/23/14 03:36]

First, sessions must occur within a short time span of each other. Given the fast-paced, instant nature of Twitter, an assumption is made that as the time span between two tweets becomes longer, those tweets are less likely to actually be related—even if there are similarities in the language used in both. For instance, one could imagine that a user might tweet about a topic such as college basketball on different days. Though the tweets might be similar based on various metrics, the user may in fact be discussing different games, and therefore those tweets likely belong to different sessions. In this study, the time limit is set at within 24 hours of the anchor tweet. The anchor tweet is that tweet to which all others within the session are related at a given threshold.

Furthermore, session length is one of the major variables measured in this study. Length is defined in two ways. First, it is understood in terms of the number of tweets within the session. At minimum, the technical definition of a session requires that there be at least two related tweets for a session to exist. Additionally, length is measured in terms of the number of characters and words contained within the tweet session. There are no actual constraints on the number of characters or words a session can contain, however, a single tweet can have at most 140 characters.

Tweet authorship and intended audience are interrelated concepts. Both deal with the exclusion of certain types of tweets in sessions. In terms of authorship, this study is most concerned with the influence of retweets versus originally authored tweets on

timelines. Retweets are defined as tweets made by one user that are reposted on another user's timeline. Though understanding the use of retweets is an interesting topic, this study is more concerned with capturing Twitter users' own voices in their tweet session. In other words, a topically cohesive session involving only retweets might be considered a 'retweet session' rather than a tweet session. Still, retweets are a major component of many Twitter timelines and can reveal a lot about a user's interests and state of mind; as such, this study did not seek to eliminate them entirely. Instead, a tweet session can contain retweets, but must contain at least one originally authored tweet as well.

Retweets are broadcast to the user's entire timeline; in this way, they are meant to be public. On the other hand, replies—tweets beginning with @username—are a form of directed communication. While others can see a public user's replies, these tweets are technically only directed to the user mentioned therein. Because of their conversational, relatively private quality, this study does not include them in sessions. They present another facet of discourse structure that the current study does not seek to address.

Signals of Cohesion

The previous qualities are easy to determine in that they deal with characteristics external to the tweet text itself. Determining the cohesion of tweets—how well they relate to each other—is a less-straightforward problem that can be approached in a number of ways. In order to be able to detect these topically cohesive tweet sessions, the meaning of 'topic' must be defined. Based on definitions found in previous research, a topic as understood in this study refers to a division of text into coherent segments (Hearst, 1994).

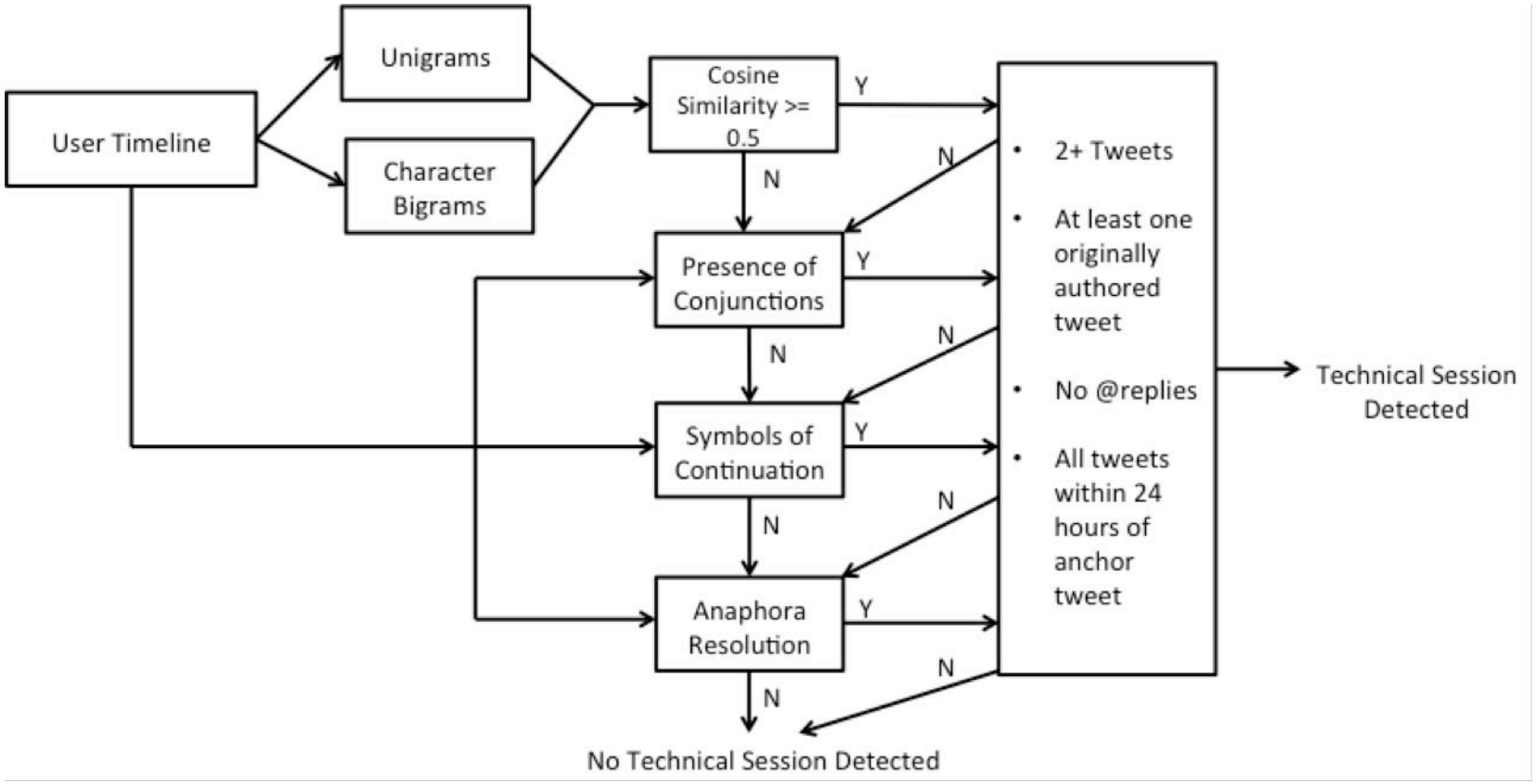
Furthermore, a change in topic is marked by changes in vocabulary; topical cohesion is marked by a similar vocabulary (Purver, 2011).

As such, in addition to the general principles of session existence, four main approaches to detecting similarity, largely based on the language within the tweets, were considered in determining whether a series of tweets were likely to constitute a session. They include repetition of unigrams and character bigrams, the use of conjunctions, the presence of terms or symbols signifying continuation, and the presence of anaphora (Table 2). Figure 1 provides a general overview of the steps taken to detect sessions in this work. I wrote Python code and performed limited manual analysis to accomplish these tasks.

Table 2: Characteristics of Technical Sessions

Characteristic	**Examples**
Repetition of strings: Presence of the same unigrams or character bigrams in two or more tweets—weighted based on term frequency-inverse document frequency (TF/IDF)—signifies a session	Unigrams My data is up.... [1/23/14 0:23] And jus like that my data is back... [1/23/14 2:27] No data..I'm done [1/23/14 23:17] Character Bigrams Bored, [1/25/14 16:47] Boredd, [1/25/14 16:48] Boreddd [1/25/14 16:48] So bored [1/25/14 17:38] --- Trying to live the right way but nobody sees that [1/19/14 14:21] When you try to do right they find everything to make it seem like you so wrong but o well [1/20/14 14:19] Got a call from a higher power just now an they told me you are going to be so blessed keep doing what you doing [1/20/14 14:22]
Conjunctive Terms: Presence of a coordinating conjunction, conjunctive adverb, or subordinating conjunction at the beginning of a tweet that connects the tweet to surrounding tweets signifies a session	I never knew what I wanted to do as a career but strangely a suitable one just fell into my lap [1/22/14 9:21] And the same way I found this career is the same way I found this job [1/22/14 9:22]
Signals of continuation: Presence of certain symbols, including '(num/X)', '+', 'lrt', at the end of tweet signifies a session with surrounding tweets	RT @(USER): if i sub or tweet about you like crazy i either care, or really like you and or want you to know some things i prolly wouldn…"[1/19/14 6:38] My LRT is so true omg.,[1/19/14 6:41]
Anaphor resolution: Presence of pronouns (he, she, it, them, they) in a tweet that refer to a noun in a surrounding tweet denotes a session	!!! "@(USER): Dude was a Boss ____O_ "@(USER): Amean..timi could dub an assignment upside down.."[1/26/14 0:31] And he was one of the smartest guys there.. [1/26/14 0:31]; Them smart children...[1/26/14 0:32]

Figure 1: Methodology Flow Chart



Repetition of unigrams and character bigrams. In order to compare texts in terms of similar language use, I utilized a Python-based library called Scikit-learn.[4] It is a library of functions for machine learning and natural language processing tasks. The main steps in this process for the purposes of this study were:

1. Vectorize the text: In order to make comparisons between texts, the documents must be in a format that can be mathematically transformed by a computer program; leaving them in their original natural language text form is insufficient. In this study, a simple "bag of words" (or, "bag of n-grams") representation is used to transform each document (here, a tweet) into a term vector. This involves tokenizing the text into its constituent n-gram parts, counting the frequency of each token, and normalizing the final frequency distribution in order to make reasonable comparisons between documents.

   In terms of tokenization, this study compared two approaches. One set of vectors was computed using unigrams (Figure 2). In these instances, tweets were split at spaces into simple one-word types. Another set of vectors was computed using character bigrams (Figure 3). Here, the tweets were split into all combinations of consecutive characters of length 2. The second approach was used specifically because it handles non-standard texts better, particularly in terms of spelling errors.[5] For instance if one word type is accidently spelled multiple ways, when frequencies are counted those tokens will be considered different types. Since character bigrams compare two characters at a time, it is much more

[4] Scikit-learn: Machine Learning in Python, http://scikit-learn.org/stable/index.html
[5] Feature Extraction, Scikit-learn, http://scikit-learn.org/stable/modules/feature_extraction.html

likely that it will be able to recognize some level of similarity between such texts—presuming the spelling errors are not too egregious.

Stopwords were removed in the process of tokenizing based on unigrams. However, stopword removal was not performed when tokenizing into character bigrams. This decision was made based on the Scikit-learn library's implementation. The stopword removal parameter had no effect when the unit of tokenization of the vectorizer was set to two characters.

Figure 2: Tokenizing tweets using unigrams

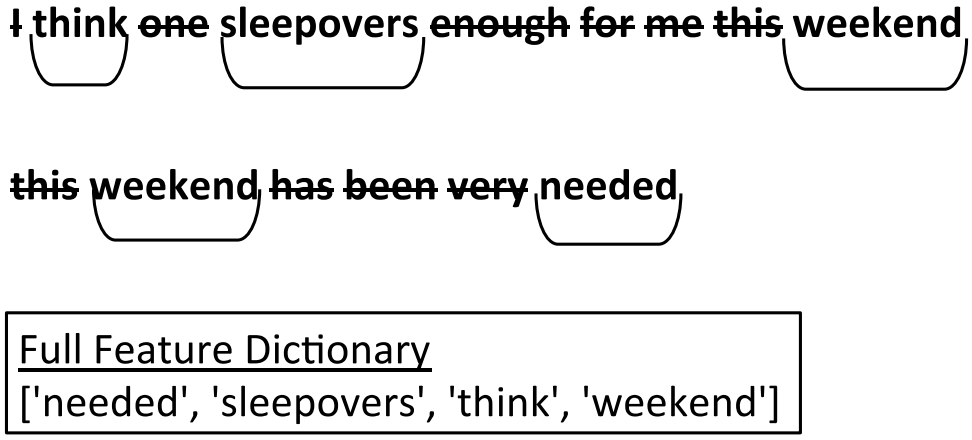
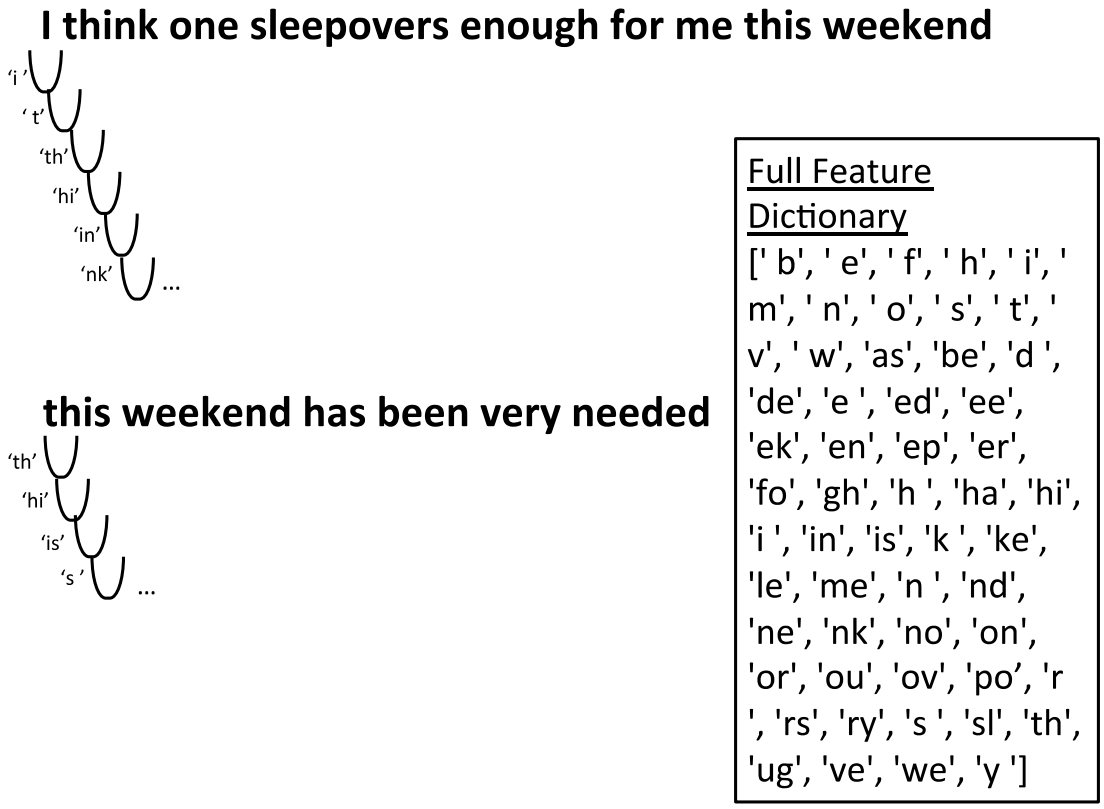


Figure 3: Tokenizing Tweets using Character Bigrams



Term frequency – inverse document frequency (TF-IDF) was used in the normalization step for both unigrams and character bigrams; it considers the frequency of tokens both in a document (i.e. a single tweet) and across the corpus (i.e. a user's timeline with replies removed). The n-grams that were most frequent in a given tweet but rare across the entire timeline were given the most weight in the representation.

2. Compute the similarity between the tweet and all other tweets on the same timeline. The prior step transforms timelines into numerical representations of each tweet. From there, it is relatively simple to compute the similarity of each

tweet based on its distance from others. One common similarity metric used in text analysis is cosine similarity.

$$k(x, y) = \frac{x\dot{y}^\top}{||x||||y||}$$

Figure 4: Cosine Similarity equation[6]

This metric is equivalent to the dot product of the vectors, divided by the product of each vector's length. Scikit-learn also has a built in function that can be used to quickly compute these values for every vector.

3. Compare the levels of similarity at a given threshold. The result of the cosine similarity measure as implemented in Scikit-learn is a series of lists, one for each tweet, with values between 0 and 1 that represent the distance between a tweet and all others. If the value is 0, there is no similarity; if the value is 1, the tweets contain the same tokens. In this study, a threshold of 0.5 was used to compare tweets. This would imply that related tweets have the majority of their n-grams in common. A script was written to return all unique combinations of tweets on each timeline where at least two tweets had cosine similarity values of 0.5 or greater.
4. Keep only those similar tweets that meet the base technical session criteria (i.e., not entirely retweets and not exceeding 24 hours of the central document). Lastly, each unique combination of lexically similar tweets returned was analyzed in order to determine if they were valid based on the general rules of sessions. If they consisted of only retweets, they were removed. Additionally, if the difference in time was greater than 24 hours between the anchor tweet (i.e. the tweet where

[6] "Pairwise metrics, Affinities and Kernels," http://scikit-learn.org/stable/modules/metrics.html

the cosine similarity was 1) and the tweets returned with it (i.e. those with cosine similarity of 0.5 or higher in comparison to the anchor), those tweets exceeding the acceptable timestamp were removed from the session.

Conjunctive terms. In order to test the idea that, in some instances, tweets may serve as clauses connected by conjunctions, a list of coordinating and subordinating conjunctions, as well as conjunctive adverbs, was gathered from Towson University Online Writing Support (Table 3). As many alternative forms of the terms as could be determined were used. For instance, 'because' is abbreviated in several ways in Internet chat; therefore, variants of this term—such as 'bc' and 'cause'—were considered. I wrote a Python script in order to look for any instance of one of these terms at the beginning of a tweet. If this condition was met, the code captured that tweet, as well as the tweet preceding it. As there were not an exceedingly high number of such instances, I manually examined these candidates in order to uncover how they were used in the tweet, and whether they served to connect the separate tweets. If they served in their conjunctive form and did connect the tweets, they signified a technical session.

Signals of continuation. Based on a preliminary examination of Twitter timelines, I gathered a set of common signals of continuation (Table 4). This list is not exhaustive, however I was unable to find any additional signals. As with conjunctions, a script located any instance of these signals in a user's tweet; subsequently, the script saved that tweet and surrounding tweets. Unlike conjunctions, the position of these symbols was not considered; they could be found anywhere within the tweet text.

Table 3: Conjunctive terms

[Conjunctions taken from Towson University Online Writing Support, "Conjunctions": http://www.towson.edu/ows/conjunctions.htm]

Type of Conjunction	Terms
Coordinating Conjunctions	for, and, nor, but, or, yet, so
Conjunctive Adverbs	after all, in addition, next, also, incidentally, nonetheless, as a result, indeed, on the contrary, besides, in fact, on the other hand, consequently, in other words, otherwise, finally, instead, still, for example, likewise, then, furthermore, meanwhile, therefore, hence, moreover, thus, however, nevertheless
Subordinating Conjunctions	after, in order (that), unless, although, insofar as, until, as, in that, when, as far as, lest, whenever, as soon as, no matter how, where, as if, now that, wherever, as though, once, whether, because, provided (that), while, before, since, why, even if, so that, even though, supposing (that), how, than, if, that, inasmuch as, though, in case (that), till

Table 4: Signals of Continuation

Type of Continuation Signal	Strings
Last Tweet/Retweet	'lrt' (last retweet) 'mlrt' (my last retweet) 'lt' (last tweet) 'tmlrt' (to my last retweet) 'pt' (previous tweet) 'prt' (previous retweet) 'continuedtweet' '#continued'
Numbering Tweets	e.g. '1/X', '2/2'
Other symbols	'+'

Anaphora Resolution. The resolution of anaphora by computational means is often a challenging task when using traditional, well-formatted texts. These difficulties are only magnified when using short, non-standard texts—such as tweets. Analysis is

predicated on determining part of speech and recognizing when various parts-of-speech (traditionally, pronouns) reference earlier parts-of-speech (often nouns). While part-of-speech taggers exist and can often perform with accuracy in the 90% range, they are generally trained on traditional media; subsequently, their accuracy on tweets is poor. As such, I identified a small subset of timelines, for which no sessions were discovered by the previous three means described, and manually analyzed them to determine if any of the tweets contained anaphoric references.

Much like the previous two tasks, if any of a set of pronouns were found in a tweet, it and the previous tweet were collected. The pronouns considered in this study were: 'you,' 'your,' 'yours,' 'he,' 'him,' 'his,' 'she,' 'her,' 'hers,' 'it,' 'its,' 'we,' 'us,' 'our,' 'they,' 'them,' 'their,' 'theirs.' First person singular pronouns were excluded as it seemed unlikely that a user would refer to himself or herself using a noun in one tweet and follow it with the pronoun 'I'. Also, as Twitter is a very individual-centered medium, excluding first-person singular pronouns drastically cut down on the number of tweets that had to be manually examined for anaphora. Future work might consider whether or not this is a valid assumption.

If the sets of tweets returned were valid based on temporal and authorship qualities, they were manually marked up. Any instance where a tweet contained a pronoun before a noun instance—and was preceded by a tweet containing a noun—was considered a possible session. If the reference in the second tweet was actually likely to refer to the noun in the previous tweet—based on agreement in terms of number and gender—it was classified as a technical session.

The ultimate purpose of this study was to find any series of tweets that met the technical definition of a session in that they fulfilled the requirements outlined above. More specifically, this research does not extend to the evaluation of such sessions based on human judgment of the actual discursive intent of these tweets. For this reason, the term "technical session" will be used throughout this study to refer to any series of tweets that met the described criteria.

Data Collection

Two main objectives dictated the data collection plan for this study. First, a set of random users was identified. Next, the timelines of this set of users was collected. Twitter has a robust set of application programming interfaces (APIs) that can serve both of these purposes. Of particular interest in this study are the GET statuses/sample and the GET statuses/user_timeline functions.[7] As stated in its documentation, the first API function "returns a small random sample of all public statuses." The second can be used to capture a public user's timeline, retrieving up to the last 200 tweets for any user specified. Both return user statuses, which consist of a wide range of metadata associated with a single post. Of primary interest here are the user id, the tweet text, the date, and the language.

Data collection took place on January 26, 2014. The first 275 user ids retrieved through GET statuses/sample were extracted for inclusion in the dataset. After removing timelines that were entirely non-English, as well as those that went private before the actual data was collected, a total of 220 user timelines were analyzed in this study. In a

[7] Twitter REST API Documentation – (1) GET statuses/sample: https://dev.twitter.com/docs/api/1.1/get/statuses/sample and (2) GET statuses/user_timeline: https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

2014 investor's report, Twitter revealed that it had an average of approximately 241 million users per month as of December 2013; 54 million of these are United States based users.[8] As such, this sample size of user accounts is equivalent to approximately 0.0001% of active monthly users (0.0004% of active US users).

For each of the 220 user ids collected, GET statuses/user_timeline was called in order to retrieve the most recent 200 posts of each user in the sample. In cases where the user did not have a total of 200 tweets, his or her entire timeline up to the date of collection was retrieved. The final sample consists of a total of 31,280 tweets. In keeping with the session definition as conceptualized in this study, all simple replies (i.e., those tweets beginning with the format '@username') were excluded from the sample; retweets were maintained. A Python script was used to remove all URLs and special characters (all those that are non ASCII alphanumeric or punctuation) from the tweets in the sample.

Ethical concerns

The ability to make one's Twitter account either public or private, and the fact that Twitter does not allow unauthorized users or applications to see and gather tweets classified as the latter, provides one easy way of making sure that non-consenting parties are not included in the study. By making one's account public, a user inevitably leaves open the ability for others to make use of the data that they publish on the social networking site.[9] Still, usernames and user ids are not identified in this study in an effort

[8] "Twitter Reports Fourth Quarter and Fiscal Year 2013 Results," 5 February 2014, https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321

[9] From Twitter Terms of Service: "By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and

to maintain privacy, following the conventions found in much of the Twitter-related academic literature.

---

distribute such Content in any and all media or distribution methods (now known or later developed). You agree that this license includes the right for Twitter to make such Content available to other companies, organizations or individuals who partner with Twitter for the syndication, broadcast, distribution or publication of such Content on other media and services, subject to our terms and conditions for such Content use." (https://twitter.com/tos)

Findings

In this section, I present the distribution of technical tweet sessions, based on lexical similarity. I discuss general characteristics of these sessions in terms of tweet count, word count, character count and retweet inclusion. For the remaining user timelines for which no session was detected based on lexical similarity, I present the number of technical sessions that were retrieved based on the presence of conjunctions, signals of continuation, and anaphora.

Lexical Similarity

Timelines were analyzed both in terms of similarity at the unigram level, and at the character bigram level. A similarity threshold of 0.5 was used for both; this implies that the majority of unigrams or character bigrams observed in one tweet were also observed in another tweet. Table 5 presents the total technical sessions detected by each tokenization technique. Additional sessions were uncovered when character bigrams were used for comparison, with 1391 sessions detected compared to 815 for the unigram approach. Similarly, slightly more user timelines were observed to have at least one technical session when using character bigrams (185 timelines) compared to unigrams (171 timelines).

Table 5: Technical Sessions Based on Lexical Similarity

	Unigrams	**Character Bigrams**
Count of User Timelines	220	220
Total User Timelines with at least 1 Session	171	185
Number of Sessions per Timeline		
0	49 (22%)	35 (16%)
1 - 5	125 (57%)	101 (46%)
6 - 10	28 (13%)	47 (21%)
> 10	18 (8%)	37 (17%)
Total Technical Sessions on all Timelines	815	1391

Word count and character count also differ when using unigrams or character bigrams as the level of tokenization (Table 6). After finding the average word count and character counts per session for each user timeline containing at least one session, I computed the overall arithmetic mean of these counts across all timelines. Again, average word count per session and average number of characters per session were both greater when using character bigrams as compared to unigrams. An average word count of 76.59 and character count of 483.35 was observed over all timelines when using the former, as compared to 35.93 and 227.31 when using the latter. The number of timelines with sessions that on average exceeded the 140-character limit imposed by Twitter was also greater when tokenizing with character bigrams—143, as compared to 90 for unigram tokenization. Conversely, the number of timelines with sessions of average character count less than or equal to 140 was greater using unigrams; a total of 81 timelines had average session size less than or equal to the size allowed in a single tweet when unigrams were used to tokenize, as compared to 42 when using character bigrams. It must be noted that these counts do not take into account URLs or special characters; as such,

they are conservative estimates of these variables. The average tweet count per session is similar when using either unigrams or character bigrams (Table 6). On average, a user's sessions contained 2.4 tweets when similarity is based on unigrams; for character bigrams, this value was slightly higher at 3.07 tweets per session.

Table 6: Estimated Average Tweet Count, Word Count and Character Count per Technical Session *

	Unigrams	**Character Bigrams**
Average Number of Tweets per Session	2.44	3.07
Average Word Count per Session	35.93	76.59
Average Number of Characters per Session	227.32	483.35
Number of Sessions > 140 Characters	90	143
Number of Sessions <= 140 Characters	81	42
* For each, the estimated average is equal to the simple arithmetic average of all averages observed on timelines with n >= 1 sessions		

Unsurprisingly, the use of retweets, which are essentially quotes, is fairly widespread among sessions as detected through lexical similarity. Table 7 provides a breakdown of the composition of sessions in relation to the number of retweets that they contained. In the table, a session containing "few" retweets is defined as one in which less than 50% of the tweets were retweets, a session containing "half" retweets is one in which 50% of the tweets were retweets, and a session containing "many" retweets is one in which more than 50% of the tweets were retweets. While the majority of sessions did not include retweets, both for unigrams and character bigrams at 0.5, they do appear in 46.1% of unigram sessions and 37.6% of character bigram sessions.

Table 7: Retweet Distribution of Technical Sessions based on Lexical Similarity

	Unigrams	**Character Bigrams**
No RTs (=0%)	439 (53.9%)	868 (62.4%)
Few RTs (<50%)	49 (6%)	116 (8.3%)
Half RTs (=50%)	273 (33.5%)	253 (18.2%)
Many RTs (>50%)	54 (6.6%)	154 (11.1%)
Total technical sessions including 1+ RTs	376 (46.1%)	523 (37.6%)

Other Approaches

Though lexical similarity detected technical sessions on the majority of timelines, there were a number of timelines for which none were observed using this method. Table 8 lists the number of timelines where sessions could not be detected based on term or character similarity. In total, there were 28 timelines for which no sessions were found using either unigrams or character bigrams. In order to determine if other methods might reveal additional relationships between a user's tweets, I analyzed the use of conjunctions, signals of continuation, and anaphora in the dataset.

Table 8: Number of timelines with no technical sessions (based on lexical similarity)

Level of Tokenization	**Total**
Unigrams	49
Character Bigrams	35
Both	28

Conjunctions. All timelines were analyzed for the presence of three types of conjunctions: coordinating conjunctions, subordinating conjunctions, and conjunctive

adverbs. Tables 9a, 9b, and 9c show the distribution of conjunction appearances at the beginning of a tweet, as well as the number of instances where this appearance is used as a conjunction connecting clauses—here, consecutive tweets. Coordinating conjunctions, particularly ‘and’ and ‘but,’ appear most often as their conjunctive part of speech, signifying technical tweet sessions 50% to 60% of the time. On the other hand, some coordinating conjunctions, like ‘so,’ occur often but rarely in their role as a conjunction (only in approximately 5% of instances). Similarly, most appearances of subordinating conjunctions do not serve in this role. The top occurring terms—‘why,’ ‘if,’ ‘when,’ ‘that,’ ‘how,’ ‘where’—appear as different parts of speech, particularly as adverbs. The same is observed for the most commonly occurring conjunctive adverbs—‘still’ and ‘finally.’ Additionally, many of these conjunctions are seen rarely or never, especially among the subordinating conjunctions and conjunctive adverbs.

The conjunctions or conjunctive adverbs that accurately denote a likely session 50% of the time or more, and that occur more than once are: ‘and’ (or ‘&’), ‘but,’ ‘or,’ ‘because’ (and variants), ‘then,’ and ‘also.’ After considering these terms and the timelines on which they begin tweets, the number of timelines without technical sessions decreases. I found sessions on ten additional timelines after identifying those with an instance of one of these conjunctions.

Table 9a: Use of Conjunctives—Coordinating Conjunctions

Conjunction	Instances where word appears at beginning of tweet	Instances where conjunction sense used within a session	Most common use of word (if not conjunction)	Example (conjunction bolded)
for	7	0 (0%)	PREP	Come support your @(USER) in their 2nd dance of the day! Hip-hops going to be awesome! JV at 5:45 Varsity at 6:15! At SCH! [1/25/14 22:17] **For** real though. They support us threw all of our games all season time to come make some noise for them! See you there!! [1/25/14 22:18]
and (&)	60	30 (50%)		I never knew what I wanted to do as a career but strangely a suitable one just fell into my lap [1/22/14 9:21] **And** the same way I found this career is the same way I found this job [1/22/14 9:22]
but	50	30 (60%)		oh so I have come to conclusion that there is really no point in even staring to do my homework [1/6/14 0:22] **but** I still should probably study....[1/6/14 0:24]
or	14	9 (64%)		I can't feel my hands [1/24/14 12:03] **Or** my face [1/24/14 12:03]
so	136	7 (5%)	ADV; Start of statement	You brought this on yourself [1/25/14 21:41] **So** no I dont feel bad for you [1/25/14 21:42]
DOES NOT APPEAR IN BEGINNING POSITION: nor, yet				

Table 9b: Use of Conjunctives—Subordinating Conjunctions

Conjunction	Instances where word appears at beginning of tweet	Instances where conjunction sense used within a session	Most common use of word (if not conjunction)	Example (conjunction bolded)
why	149	0 (0%)	ADV	
if	144	0 (0%)		
when	85	0 (0%)	Both ADV & CONJ	
that	69	0 (0%)	PRON	
how	66	0 (0%)	ADV	
where	20	0 (0%)	ADV	
because (bc, b/c, cos, cuz, cause)	11	7 (63%)		so the other day an anon asked me to record myself singing but i need my brother's really good mic so i can actually sing [1/26/14 14:11] **because** my laptop microphone is extremely low quality i don't even understand how my skype contacts could maintain a call with me[1/26/14 14:12]
since	9	1 (11%)		big bang theory till i ko [1/22/14 0:00] **since** i can't catch this celtics game [1/22/14 0:01]
after	4	1 (25%)		tomorrow i do though, but it's 4 my last two finals [1/23/14 12:14] **after** that it's new classes oh lordy [1/23/14 12:14]

Table 9b: Use of Conjunctives—Subordinating Conjunctions

Conjunction	Instances where word appears at beginning of tweet	Instances where conjunction sense used within a session	Most common use of word (if not conjunction)	Example (conjunction bolded)
as	4	1 (25%)		i am currently addressing tradespeople wearing a repro victorian nightdress with an apron over the top.[1/24/14 16:16] **as** you do.[1/24/14 16:16]
whenever	4	0 (0%)		
as soon as	3	0 (0%)		
once	2	0 (0%)		
unless	1	1 (100%)		i want highpoly, nicely textured mmd stages [1/25/14 21:18] **bc** i can't find enough of those [1/25/14 21:18] **unless** i go on nebusoku's blog or i search through nicoseiga [1/25/14 21:18]
until	1	0 (0%)		
no matter how	1	0 (0%)		
now that	1	0 (0%)		
while	1	0 (0%)		
before	1	0 (0%)		
even if	1	0 (0%)		

Table 9b: Use of Conjunctives—Subordinating Conjunctions

Conjunction	Instances where word appears at beginning of tweet	Instances where conjunction sense used within a session	Most common use of word (if not conjunction)	Example (conjunction bolded)
even though	1	1 (100%)		can friday hurry up already [9/23/13 20:50] **even** though the week just started [9/23/13 20:51]
DOES NOT APPEAR IN BEGINNING POSITION: in order (that), although, insofar as, in that, as far as, lest, as if, wherever, as though, whether, provided (that), so that, supposing (that), than, inasmuch as, though, in case (that), till				

Table 9c: Use of Conjunctives—Conjunctive Adverbs

Conjunction	Instances where word appears at beginning of tweet	Instances where conjunction sense used within a session	Most common use of word (if not conjunction)	Example (conjunction bolded)
still	26	0 (0%)	Simple ADV	
finally	19	0 (0%)	Simple ADV	
then	8	5 (62.5%)		everyone always leaves.[1/24/14 18:40] **then** they wonder why i run..[1/24/14 18:40]
next	7	0 (0%)	ADJ	
instead	4	1 (25%)		liam should just admit he fucked up and apologize instead of making himself the victim and trying to cover it up, as simple as that [1/19/14 14:44] **instead** of having his sister say the things that he should have said [1/19/14 14:44]
also	2	2 (100%)		i can't imagine that talib means that much to the pats. if so then what a joke. the guy is scum bag and karma is a bitch for him #adderall [1/19/14 22:27] **also** denver has settled for way too many fg's. better teams will make u pay for that. [1/19/14 22:30]
meanwhile	1	1 (100%)		i can't sleep but i gotta take a nap…[1/2/14 19:50] **meanwhile** i'm chillin [1/2/14 19:51]
DOES NOT APPEAR IN BEGINNING POSITION: after all, in addition, incidentally, nonetheless, as a result, indeed, on the contrary, besides, in fact, on the other hand, consequently, in other words, otherwise, for example, likewise, furthermore, therefore, hence, moreover, thus, however, nevertheless				

Signals of Continuation. A small set of markers that denote continuation was identified in the corpus. Their frequency across user timelines is presented in Table 10. Several of the hypothesized features were not found in the sample. The features that do appear are text strings that explicitly reference an earlier tweet. Though all observed instances were cases where a session was observed, these signals were relatively rare, occurring only 13 times across the entire dataset. Sessions for one additional timeline for which no technical sessions had been discovered by other means were found. That timeline contained two separate technical sessions, and each of these sessions contained the last tweet/retweet string, 'lrt.'

Table 10: Signals of Continuation

Signals	**Instances Observed**	**Instances Used in a Session**	**Example (signal bolded)**
Last Tweet/Retweet Abbreviations	13	13 (100%)	RT @(USER): The anthem (URL) [1/26/14 0:57] My theme song my **lrt** [1/26/14 0:58]
DOES NOT APPEAR: Numbering tweets [i.e. (1/X)], '+'			

Anaphora Resolution. Lastly, timelines for which no sessions could be detected by other means were examined for the presence of anaphora. In total, this 'stubborn set' consisted of 19 timelines. Table 11 presents the distribution of potential anaphoric references and the number of instances where these resolved in such a way as to make cohesion between tweets probable. Technical sessions were observed on five additional timelines as a result of the presence of anaphoric references. This constitutes 26.32% of the remaining sessionless accounts. However, the majority of cases where a pronoun

closely followed some noun in separate tweets were not, in fact, actual sessions; only 17.5% of possible instances were technical sessions (based on anaphora). On eight of the sessionless accounts, no possible instances (based on anaphora) were observed.

Table 11: Distribution of Anaphora Analysis

Sessionless Account	Possible Sessions *	Technical Sessions **	Other Sessions Present? ***
1	8	1	Yes
2	6	0	Yes
3	6	1	Yes
4	5	0	
5	4	2	
6	3	2	
7	3	0	
8	2	1	
9	1	0	
10	1	0	
11	1	0	
12	0	0	
13	0	0	
14	0	0	
15	0	0	
16	0	0	
17	0	0	
18	0	0	
19	0	0	Yes

* All conditions (in terms of time, not all RTs, no replies) met, first tweet of candidate contains a noun, and following tweet(s) has a pronoun preceding any noun instances
** Where anaphora contribute to tweet similarity
*** Where similarity is not the result of anaphora resolution (Detected based on manual examination and assessment)

Anaphora resolution does not reveal all remaining ‘stubborn set’ sessions, but further manual analysis of these accounts shows that other sessions do still exist. These sessions went undetected based on all computational approaches performed in this study. Table 12 presents additional sessions that were detected through manual analysis from four ‘stubborn set’ timelines. As their cosine similarity measures show, most of these

sessions fell well below the 0.5 threshold; this was particularly the case when unigrams were used as the unit of tokenization. The time between tweets for these sessions ranges from as little as one minute up to approximately three hours.

Table 12: Examples of Undetected Sessions

Session	Tweet Date/Time	Tweet	Cosine Similarity (Unigrams)	Cosine Similarity (Character Bigrams)
1[10]	1/24/14 16:52	Starving children commercials always make me sad, but had their parents gone against their religion and used contraception thered be no kid	0.087	0.399
	1/24/14 16:53	I mean, take one pill a day or watch your children die from malnutrition? Is it really even a hard decision?		
2	12/24/13 6:06	All I see is retweets of naked girls. #likeno #stop	0	0.112
	12/24/13 6:20	Alright now they are doing it on purpose. @darbibragg		
3	1/7/14 4:46	Win or lose, with my brother and dad as my witnesses, I called that kickoff return.	0	0.227
	1/7/14 5:11	RT @TheTweetOfGod: Sorry Auburn. Your miracle quota was filled.		

[10] Though there is an anaphora-like reference in the second tweet connected to the first ('your' to 'their parents'), since it was not in the format considered for analysis (i.e. pronoun in the subsequent tweet(s) coming before any noun in that tweet), it was not considered resolved through anaphor use. This is a more conservative approach but more computationally simple to be programmed.

Session	Tweet Date/Time	Tweet	Cosine Similarity (Unigrams)	Cosine Similarity (Character Bigrams)
4	1/25/14 21:46	Shout out to @WAGERFUT14PS3 for being legit	0.183	0.369
	1/26/14 0:50	No one play @WAGERFUT14PS3 he's a scammer and a sore loser who won't pay up		

In several cases, however, sessions simply do not exist. Table 13 presents an example timeline for which no technical sessions were observed after all approaches were attempted. Beyond a lack of topical cohesion as observed by low cosine similarity scores, some of the issues observed on these timelines included too much time between tweets and a lack of originally authored tweets. Additionally, certain timelines primarily contained links and only a few original tweets. Since URLs were not utilized in this study as a contributor toward the measures of similarity, timelines that were largely link driven—just as with those that were mostly retweet driven—were unlikely to have sessions.

Table 13: Example Sessionless Timeline

	Tweet	Date
1	I'll be remembering the fallen at 11 o'clock #2MinuteSilence #LestWeForget	11/11/12 11:00
2	Why are there no pigeons? - new #blipfoto entry	3/13/13 16:54
3	Fighting against the winter - new #blipfoto entry	3/14/13 20:41
4	RT @keirshiels: Joyously British genuine Golf Rule Amendments for WW2. Courtesy of @FreddieVonberg HT @qikipedia	4/5/13 16:44
5	Streatham Hill Train Care - new #blipfoto entry	5/30/13 20:14
6	A car with attitude in SW16 - new #blipfoto entry	6/2/13 22:04

Ultimately, technical sessions, as detected by any of the methods outlined for use in this study, were found on 206 user timelines (93.6%). Only 14 timelines (6.4%) had no technical sessions that could be detected based on the definitions and techniques utilized in this study.

Discussion

This section discusses each of the study's research questions in relation to the observed findings. I present possible explanations for these findings, making some comparisons to other findings in the relevant literature. Lastly, this section reveals some limitations of this study.

The findings of this study in relation to RQ1— "What percentage of users engages in tweet sessions?"—suggest that sessions, as defined in this study, are found on the majority of user timelines. Of the 220 user timelines analyzed during the course of this study, only 14 (6.4%) were found to contain no sessions meeting the technical definition. In other words, 206 (93.6%) of timelines did contain some series of tweets, at least one of which was originally authored by the user, all posted within 24 hours of the anchor that were lexically cohesive, contained discriminative conjunctions, contained signals of continuation, or had anaphora that could be resolved to a reference in an earlier tweet.

In relation to RQ2—"How many sessions are detected using various techniques for detecting cohesion? How do different feature representations perform?"—lexical similarity as determined by cosine similarity between vectorized tweets was shown to have the best potential of revealing technical sessions on user timelines meeting the requirements imposed in this study. Similar tweets posted within 24 hours of each other

were discovered on the vast majority of timelines, using either unigrams or character bigrams. Utilization of character bigrams is the more liberal approach, finding prospective sessions on almost 84% of timelines in the sample, compared to 78% using unigrams. Character bigrams produce more false positives, however, as similar character strings can be found in completely different words. This is a trade-off that may be acceptable in certain applications, but unacceptable in others. If recall is more important, and human analysis will follow, using character bigrams is the more useful approach. However, if precision is more important and little to no human follow-up analysis will occur after computational analysis, unigrams are likely to be better in terms of lexical cohesion.

While lexical similarity appears to be a very promising approach, the other forms of analysis seem to have more limited utility and narrower scope. Though they were useful for revealing some sessions—particularly some of the coordinating conjunctions and the signals of continuation—there were many more instances where these techniques clearly resulted in a high frequency of false positives. For instance, subordinating conjunctions (other than 'because') seem to have little to no utility in detecting sessions. Many of the features also appeared with such low frequency as to probably not be significant.

A major factor for the lack of detection of some sessions seems to be the fact that, although they may not contain the exact same vocabulary, they tend to contain words that are related to each other or are reliant on more context to resolve references. As such, it is highly possible that a bag of words representation is a limiting approach. Augmentation

of the feature representation with synonyms as well as broader and narrower terms could lead to more sessions or better session boundaries.

With regards to RQ3—"On average, how long is a tweet session in terms of textual length?"—though average tweet session length is greater than 140 characters (or the single tweet character limit) for more than half of timelines when using either feature representation, it is interesting to note that there are still a number of sessions that could, in reality, fit into the constraints of a single tweet. This was especially true when using unigrams to compare tweets. While it still means that detected sessions can provide more information than a shorter, single tweet—two or more tweets will, by default, be longer than a single tweet—this finding is somewhat disappointing if one hopes that discovering tweet sessions will lead to a much richer document representation for use by analytical tools. If a session is little more than the maximum length of a single tweet, it is still an impoverished representation. It does, however, seem promising that a character bigram representation results in longer sessions; further augmentation could also lead to improvement. Ultimately, this confirms the trend observed in Alis and Lim (2013); tweet sessions, like tweets, are inherently short.

This finding regarding session length would also seem to suggest that the restrictions imposed by Twitter are not the only reason why people might engage in tweet sessions; in other words, there are a number of instances where users feel the need to divide up cohesive texts, not because of the constraints imposed by Twitter but based on personal choice. There are a number of other possible explanations for this behavior. For instance, many users use Twitter to share thoughts, information, and opinions as they

come to them; such users may not take a lot of time to plan out exactly what they want to say in its full form, leading them to share it bit by bit as it evolves in their own mind. Similarly, because some sessions align with events, which have a strong temporal component, users may tweet something as it happens; this may result in short, steady updates. Also, some tweets could be simple responses to other tweets, especially when a retweet is also in the session. As observed, in answer to RQ4— "How influential are retweets in tweet sessions?"—retweets show up in over one-third of technical sessions as detected through lexical cohesion using unigrams or character bigrams, with its effect being particularly notable when unigrams are used as the unit of tokenization. These may not require as much text, leading to shorter sessions.

Furthermore, there are several common discourse structures that can inform the behavior observed on Twitter timelines in relation to sessions. For instance, users may spread information across short tweets as a means to emphasize a point. This is particularly true given that lexical analysis (especially when based on unigrams) reveals repetition of terms; repetition is a common way to emphasize something, which may factor into how users tweet during a session. Distribution of text across separate tweets can similarly serve in the same way that a pause does in speech. Again, such practice would allow the user to express an idea to followers that would likely come across differently had it been simply written as a single tweet. For instance, a Twitter 'pause' could be used to humorous effect or to heighten the drama and suspense in recounting a story. Numerous studies, including Galley et al. (2003), have considered the role of pauses and silence in segmenting and understanding discourses from spoken, transcribed text. It is possible that Twitter 'pauses,' taking the form of distributed text across tweets,

also serve in some meaningful way that is useful in understanding discourse on this medium. Whether any of these possibilities are the express intent of users cannot be determined based on this research, however they are issues worth exploring in future work.

Limitations

There are some limitations to this study. Due to its exploratory nature, several of the design choices made were rather arbitrary. It is possible, and perhaps likely, that a shorter timespan may be more revelatory for accurately recognizing sessions rather than the 24-hour span enforced here. Additionally, a cosine similarity threshold of greater than or less than 0.5 may be a better choice in terms of determining lexical cohesion in session. This threshold was determined after an informal look at the qualities of the sessions—primarily the number of sessions and how likely they seemed to be related—returned at lower or higher levels, but a more systematic approach would be better. Both of these speak to the largest limitation of this exploratory study: it will be important to evaluate the performance of these metrics and threshold levels more stringently. Adding additional coders to determine actual session boundaries on timelines, discovering how often they agree, and comparing their results to the results returned using these parameters would be an important future step to better tune these values.

There also may be certain terms and term variants that were not considered in the conjunction identification, signals of continuation, and anaphora resolution steps. Given the wide variability of informal web texts, there are likely to be a number of terms that were simply not considered here. Furthermore, some of text processing may have had an

influence on ultimate performance. Particularly, excluding URLs from the final tweets and discarding special characters might cause some sessions to be missed.

Since only the most recent 200 tweets on a user's timeline were collected it is possible that some parts of the tweet session that might come before or after the point of collection may be missed. This could make the findings slightly misleading; it would be possible that a session is actually longer than the findings show (both in terms of text length and temporal length). However, this is unavoidable, as an unlimited amount of data over an unlimited time frame could not be collected.

Conclusions & Summary

Twitter has proven to have great potential in providing people and organizations with useful information that can aid the decision-making process. However, the fact that tweets are so short—at most, 140 characters—can lead to problems when processes such as sentiment analysis are attempted. This study aims to alleviate ambiguity by introducing and measuring the concept of the tweet session—instances where users express a thought or opinion on a topic using more than one tweet. Based on the definition of a technical session and the approaches used to measure similarity, it was found that almost all user timelines contain at least one session. In this sample, at least one technical session was detected on all but 14 timelines (93.6% of timelines) using the techniques outlined—and several of the remaining timelines had undetected sessions, based on a cursory manual analysis.

Ultimately, this work suggests that the tweet session is a real concept that can be discovered and measured by several computational means, making it worthy of further study. Although this research seems to suggest that many tweet sessions are still very short, it nonetheless shows that, in many cases, a very impoverished span of text can be augmented to a much less impoverished form. It is hoped that these insights will lead to future work that can help to improve analytics services using big data sources of short, noisy social data.

However, there is still much that can be done to improve detection and prove the usefulness of sessions in these applications. Some future directions include augmenting the feature representation, which may improve performance. Expanding tweets using a source such as WordNet or FrameNet would be one reasonable approach. Using trigram character strings may also be worth exploring. Also, more parameter tuning should be done to ensure that the temporal and similarity levels are at the most effective level. The temporal aspect, particularly, is one concept that seems to have a lot of potential in session detection. Deeper analysis into any tweets that occur within a very short span (such as a minute or less) could reveal interesting findings regarding how influential time is in comparison to lexical similarity in detecting sessions. Lastly, this study does not go so far as to explore user intent based on the discourse structures observed; further research should be performed to reveal more about the nature of discourse as observed on Twitter.

References

Alis, C. M., & Lim, M. T. (2013). Spatio-temporal variation of conversational utterances on Twitter. *PLoS ONE*, *8*(10), e77793. doi:10.1371/journal.pone.0077793

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1* (pp. 562–569). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1075096.1075167

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Hamilton, H., Schiffrin, D., & Tannen, D. (2001). *The handbook of discourse analysis*. Malden, Mass.: Blackwell Publishers.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information Processing & Management*, *38*(5), 727–742. doi:10.1016/S0306-4573(01)00060-7

Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 9–16). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981732.981734

Hirschberg, J., & Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3), 501–530.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (pp. 56–65). New York, NY, USA: ACM. doi:10.1145/1348549.1348556

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology*, *58*(6), 862–871. doi:10.1002/asi.20564

Jones, R., & Klinkner, K. L. (2008). Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 699–708). New York, NY, USA: ACM. doi:10.1145/1458082.1458176

Lim, E., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Trans. Manage. Inf. Syst.*, *3*(4), 17:1–17:10. doi:10.1145/2407740.2407741

Morris, J., & Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations As an Indicator of the Structure of Text. *Computational Linguistics*, *17*(1), 21–48.

Purver, M. (2011). Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, 291–317.